DOCUMENT RESUME

ED 230 586 TM 830 350

AUTHOR Wiersma, William; And Others
TITLE Assessment of Teacher Performance: Constructs of Teacher Competencies Based on Factor Analysis of Observation Data.
PUB DATE Apr 83
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classroom Observation Techniques; *Competence; Competency Based Teacher Education; Factor Analysis; *Factor Structure; Higher Education; Measurement Techniques; Performance Factors; Student Teachers; Teacher Behavior; *Teacher Evaluation; Test Validity
IDENTIFIERS *Classroom Observations Keyed Effectiveness Res; *Teacher Performance Assessment Instruments

ABSTRACT

This study focused on a comparison of the factor structures of the scores on 18 teacher competencies measured by both low-inference and high-inference observation. The observation inventories used were the Classroom Observations Keyed for Effectiveness Research (COKER) for low-inference observation, and the Teacher Performance Assessment Instruments (TPAI) for high-inference observation. The data were obtained on over 250 student teachers observed in their concluding year of training, in a four-year, competency-based, teacher education program. The factor structures based on the two sets of scores, one from the COKER, the other from the TPAI, were quite different; although variance accounted for was about the same. The TPAI produced a strong, general factor and a second factor concerned with instructional variety. The COKER produced five, consistently strong factors concerned with variety and flexibility, modifying instruction and working with various sized groups, discipline and communication, questioning and discussion, and feedback and correction of cognitive performance. The results indicated little convergent validity between the low-inference and high-inference observation. Measurement in this area appears to be highly instrument dependent. (Author/CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. W. Crsma

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED230586

ASSESSMENT OF TEACHER PERFORMANCE: CONSTRUCTS OF TEACHER COMPETENCIES BASED ON FACTOR ANALYSIS OF OBSERVATION DATA[1]

William Wiersma, George E. Dickson, Stephen Jurs and Judith Wenig
The University of Toledo

Introduction

The development and use of high and low inference observation of teacher behavior in the classroom as well as other procedures to measure teacher competence, performance and effectiveness continue to be a major concern of teacher educators, other education professionals and the public. There continues to be an increasing demand for the use of teacher assessment procedures as witnessed by the growing number of state education agencies and others in the development and use of assessment procedures.

Resistance to research, evaluation, or assessment of teacher competence continues to come from various segments of the education profession and other societal groups, usually on the argument that measurement procedures are improper and/or the results inconclusive. However, regardless of such criticism, the national drive for educational accountability through demonstrated teacher competence continues and requires constant attention as to whether teacher training programs purporting to develop teaching competencies actually prepare persons who possess the repertoire of professional knowledge and skills considered necessary for successful teaching.

Several observation instruments have been developed over the past 20 years or so, designed to measure teacher-pupil interaction in the classroom, or to measure teacher behavior. Observation may take various forms

TM 830 350

---

[1]Paper presented at the 1983 Annual meeting of the American Educational Research Association, Montreal, Quebec, Canada; April 1983.

and the periods of observation vary considerably across instruments. The study reported in this paper involved the observation of student teachers using two observation inventories; the inventories being quite different in approach. Before discussion of analysis and results, some information is provided about the specific inventories and the process of their use.

Low-Inference and High-Inference Observation

The recording of data from classroom observation has taken various forms, running the continuum from tabulations in matrices representing classroom interaction to recording anecdotal data an subjectively perceived by the observer. More recently, the terminology has become low-inference and high-inference, as characterized by the amount of inference required from the observers. Some writers would argue that the extent of inference is on a continuum, and there may be some merit to this argument. For the purposes of this research, we will maintain the low-inference, high-inference dichotomy, designating any observation that requires the observer to make a qualitative judgment of the classroom interaction as high-inference observation. Observation requiring only a tabulation of defined behaviors is considered low-inference observation.

The advocates of low-inference observation and those for high-inference observation provide convincing arguments for their respective approaches. Advantages cited by one are often viewed as disadvantages by the other. Undoubtedly both approaches have their merits and weaknesses. Presently, indications are that the jury is still out as to the better approach. The method of observation is a very important matter, however. In the final analysis, it is the way measurement is conducted that operationally defines whatever is measured by the observation.

Low-inference observation generally involves recording the occurence of specific behaviors which are defined with sufficient precision so that

observers will agree of whether or not they have occured. It is a behavioristic approach throughout. To know what occurs in the classroom, it is necessary to observe the teacher and students. Low-inference observation does not deal with nonclassroom, supportive activities such as writing lesson plans. The assumption is that such activities by themselves do not impact directly on student learning. In order for a behavior to have an effect it must be demonstrated in the instructional context.

Low-inference observation has an advantage in its objectivity. There may be some subjectivity involved in placing an observed behavior in a matrix, but generally this is relatively minor when using trained observers, as must be the case with low-inference observation. There is no option for putting a qualitative judgment on the behavior.

Low-inference observation is generally more straightforward than high-inference observation and it considers behaviors in more basic components. Nonclassroom behaviors which may be supportive of instruction are not observed. It avoids the possibility of erroneous inferences about high quality supportive behaviors necessarily resulting in effective teaching.

Even with its objectivity, low-inference observation has some disadvantages. Opponents would challenge its limited scope by focusing on certain classroom behaviors of teacher and students. This disagreement is probably more a disagreement in philosophy about teaching, the generality-specificity issue in teaching, or perception of the nature of teaching and learning. Nevertheless, in view of the present lack of consensus among research results on effective teaching, it is important to retain the option of measuring what the profession (or segments thereof) perceives to be important.

High-inference observation in classrooms requires an observer to make substantial inferences as to what teacher behaviors are being demonstrated and the quality of such demonstration. As with low-inference observation,

data can be obtained for competencies or groups of behaviors. Observers, usually working from a set of indicators and descriptors for the competencies (or whatever is being observed) make judgments on particular aspects of teacher performance.

The concept of including a qualitative judgment about teacher behavior is intuitively appealing, and it has a practical advantage in the scoring procedure over low-inference observation. Scores can be generated on a scale of measurement that at least approaches an interval scale. Scores can be considered in an absolute sense and comparisons can be made across competencies.

The major criticism of high-inference observation centers on the subjectivity which can enter in various ways. The subjectivity of the quality of what is being observed is the most obvious source. Opponents of high-inference observation often suggest that an individual's score is more a function of who does the observing than what actually occurred. This criticism may be countered, at least in part, by a consideration of professional expertise. Consider an example from the medical profession. Two medical students may both be competent to make incisions and to put in sutures for a defined surgery. Yet the quality of the surgery performed may differ, and we generally accept the notion that a trained surgeon can distinguish the quality. It is similarly claimed in education that trained observers should be able to distinguish quality when observing the teaching-learning situation.

Another factor of subjectivity is introduced if an absolute "standard" is imposed for judging performance, such as a mean of 3.0 (using the 1-5 scale) across the behaviors for a competency. Certainly, this would be a subjective judgment. Presently, there is little if any empirical evidence by which standards can be set.

The low-inference, observation inventory used in this study was the Classroom Observations Keyed for Effectiveness Research (COKER). The COKER is a relatively comprehensive system for recording pupil and teacher behavior as it occurs in the classroom. The inventory is applicable for both elementary and secondary level classes.

The COKER consists of a matrix of 389 cells in two sections: Section A is 2-dimensional and represents the interaction between teacher and students; Section B is 1-dimensional describing the teacher's and student behaviors. The COKER is the result of an extensive developmental process and is a composite instrument derived from five objectives, systematic, direct observation instruments used in studies of teaching.[2] The items on the COKER have been generated and adapted from more than 1,300 category sign items on these five different observation instruments. The intent of synthesizing items from the five instruments was to capture, to the extent possible, the breadth of the original instruments. Selection of items was based on correlations between scores on individual items and pupil gain and the initial inventory was called the Georgia Assessment of Teacher Effectiveness (GATE) which became with further, slight modifications, the COKER. The COKER is likely the most comprehensive and useful instrument available for low-inference observations of teacher behavior.

The high-inference instrument used for this study was the Teacher Performance Assessment Instruments (TPAI). This instrument from the University of Georgia was developed through an extensive review and consensus process. It was created in the context of the competency-based teacher education program at the University of Georgia and the decision of the Georgia Department of Education to employ teacher competence considerations in teacher

---

[2]These instruments were the Observation Schedule and Record, Form 5 Verbal (OSCAR), Spaulding Teacher Activity Rating Schedule (STARS), Florida Climate and Control System (FLACCS), Teacher Practices Observation Record (TPOR) and Coping Analysis Schedule for Educational Settings (CASES).

certification. Hundreds of teaching competency statements from a variety of sources were collected, sorted and synthesized. A survey instrument consisting of 52 competency statements was administered to a stratified sample of 4,668 Georgia educators. Only competencies consistently classified as generic and essential were retained resulting in 22 competency statements for inservice teachers and 17 for student teachers. The TPAI is now used in Georgia, providing evidence through observation for certification.

The TPAI contains five categories of competency statements called instruments. They are:

Teaching Plans and Materials (TPM)

Classroom Procedures (CP)

Interpersonal Skills (IS)

Professional Standards (PS)

Student Perceptions (SP)

The final category is not contained in the student teacher version of the TPAI. In fact, this category is not part of any observation but consists of a 30-item paper and pencil rating scale to which a teacher's pupils respond.

The four categories requiring observation included 51 teacher behaviors called "indicators." Each indicator has a set of descriptors designed to generate a score from 1 to 5. The greater the number of the descriptor the more complete the teacher's behavior on the indicator.

The Measurement of Competencies

How do we get from occurrence of behavior as with low-inference observation, and scores of 1 to 5 on indicators as with high-inference observation, to scores on teacher competencies? A prior question is, "Where do we get teacher competencies?" The processes used in arriving at answers to these questions are by no means trivial. They are briefly

described below in the context of The University of Toledo, competency-based, teacher education (CBTE) program.

The intention of the study was to measure the competencies underlying the CBTE program, and those competencies were identified through a thorough content analysis of program instructional modules. The analysis was conducted by three faculty members familiar with the CBTE program. All modules were reviewed, and the above procedure produced 49 competencies common to both elementary and secondary levels. An example of a competency is, "Establishes transitions and sequences in instruction which are varied."

However a competency is stated, it is operationalized by the observation instruments used to measure it. The decision to use the COKER and TPAI was made after an extensive review of available inventories. Generating competency scores requires a mapping of the behaviors measured by the inventories onto the competencies. Mapping is a task of synthesis, successive approximations, and operationalizing descriptive terms in the context of the inventories. Essentially, the mapping process results in identifying the behaviors that contribute to a competency (represented by cells in the COKER and indicators in the TPAI). However, the process also contributes to the formulation of competencies. If for example, two competencies consist of the same or very nearly the same behaviors, the two competencies are identical as far as measurement is concerned, regardless of what the competencies were called initially.

Of the 49 competencies identified for the CBTE program, 45 were measured by the TPAI and 22 by the COKER. Generally, low-inference inventories tend to measure fewer competencies because they measure only competencies demonstrated in the classroom; they do not measure supportive competencies

in areas such as planning. The focus of this paper is upon the 18 competencies measured by both the COKER and the TPAI.

Scoring TPAI and COKER Data

Before considering results, a few comments about scoring observation data are in order. Generating a score for a competency using the TPAI is quite straightforward since the TPAI provides scaled scores on the indicators. There are varying numbers of indicators per competency, so to determine a score on a competency, the indicator scores are summed and divided by the number of indicators. Thus, the TPAI score for a competency is the mean score of the indicators that map onto the competency. Scores may range from 1 to 5 inclusive. The indicators that contribute to a competency are weighted equally.

The COKER data do not contain a qualitative characteristic, so scaled scores are not directly obtained from the data as with the TPAI. The data of the COKER consist of check marks indicating the presence of behavior. These data need to be converted to a competency score. Scores based on tallies or frequencies of items checked are not adequate to accommodate the varying numbers of items per competency, or varying numbers of observations. So, some type of standardization of scores or proportionality is required in the scoring. Also, COKER items may be scored on presence or absence of behavior, and that characteristic needs to be accommodated.

There are different, possible approaches to scoring low-inference, observation data, and the approach used for the Toledo data is called the "Proportionality Score Model." This approach to scoring COKER data provides a value (p) for each item or cell that maps onto a competency, which is the proportion of times the item was checked over however many times the teacher was observed. For items scored on absence of behavior, 1-p rather

than p is used. These proportions are then summed across the items on the competency. Since competency scores are based on varying numbers of items, this sum is divided by the number of items in the competency, so that scores can be compared across competencies. Scores so generated can take on values from zero to 1.0 inclusive. These are multiplied by 100 simply to provide a more familiar scale.

Reliability of Observation Inventories

Instrument reliability often involves problems when dealing with observation, not only from an empirical standpoint, but also from a theoretical perspective. Both the COKER and the TPAI require the use of trained observers. Observer agreement is important, but it is not the reliability of a competency score.

The developers of the TPAI have conducted numerous reliability studies, both with beginning teachers and student teachers. Capie, et al (1979)[3] provided a summary of results. Cronbach Alpha internal consistency coefficients were computed for the ratings on each of the three major instruments making up the TPAI. A number of raters, varying from 11 to 16, provided ratings in third grade science and eighth grade language arts. The Cronbach Alpha coefficients ranged from .65 to .80 for third grade science and .44 to .75 for eighth grade language arts across the categories.

Reliability information, conceptual and empirical, is provided in the Users Manual (1982)[4] for the COKER, and the details will not be repeated

---

[3]William Capie, ET AL, SELECTED INVESTIGATIONS OF THE RELIABILITY OF THE TPAI. (Athens, Ga.: University of Georgia working Paper, Teacher Assessment Project, College of Education, The University of Georgia, 1979).

[4]Joan G. Coker and Homer Coker. CLASSROOM OBSERVATIONS KEYED FOR EFFECTIVENESS RESEARCH, USERS MANUAL. (Atlanta, Ga.: Georgia State University/Carroll County Teacher Corps Project, 1982.) pp. 13-19, 31-37.

here. The reliability of greatest concern is the internal consistency reliability estimated through the Cronbach Alpha coefficient. Essentially this is the internal consistency of the items which apply to a competency. Stability reliability may also be considered; however, maintaining stability across a series of observations may be of little concern, especially when observing student teachers. They may be teaching different subjects, especially in the elementary grades, or there may be direct attempts to change teaching performance.

Estimation of internal consistency coefficients has some problems associated with it. For example, on occasion a competency may have only one item associated with it, so the α-coefficient does not apply. Coefficients were as high as .80 or so, but generally they tended to be more modest; in the .40s and .50s. Competencies with small numbers of items tend to have the lowest reliability coefficients, although this characteristic is by no means without exception.

Data Collection For the Study

Data were collected over a period of six academic quarters involving a total of 153 student teachers. Each student teacher was observed 16 times with the COKER and twice with the TPAI. The COKER observation period is five minutes, with another five minutes used for completing Section B from memory. A second observation follows immediately during the same class period. A TPAI observation period is the length of a class period, usually around 50 minutes.

The student teachers observed were preparing for elementary and secondary levels, with some student teachers in special education. A variety of subjects being taught was observed, both at the elementary and secondary levels. Student teachers were informed that they would be observed, but no

special arrangements were made. Results were used for program evaluation and research purposes, and the student teachers were told that they would not be evaluated individually on the observation results.

The focus of the present analysis is the set of 18 competencies that is measured by both the COKER and the TPAI. These 18 competencies are listed in Table 1. It was hypothesized that a student teacher who was judged to be successful on one instrument would alse be judged to be successful on the other instrument. The convergent validity of the two instruments would support the construct validity of each.

Factors that could mitigate against the hypothesized correlations include the fact that different observers used the different instruments and that the observations did not take place on the same day. Hence there is the strong assumption that the student teachers would be consistent in their behaviors and that assumption was not specifically tested. However, the two methods for measuring the same teaching competencies of the same student teachers would be correlated.

A related interest that was served by the analysis was the factor structure of the set of competencies that was measured by both the COKER and the TPAI. The numerous competencies appear to be separate skills but in fact they may measure fewer dimensions than the 18 separate competencies In the extreme, one would expect the factor structure to yield 18 factors with two measures loading significantly on each factor, one measure being a competency measured by the COKER and the other measure being some competency measured by the TPAI. Of course a more realistic expectation would be to have several of the competencies merge into some general factors of teaching.

Table 1

List of Competencies Measured
by Both the COKER and the TPAI

Competency	Description
1	Uses a variety of instructional strategies.
2	Establishes transitions and sequences in instruction which are varied.
3	Modifies instructional activities to accommodate identified learner needs.
4	Demonstrates ability to work with individuals, small groups and large groups.
5	Structures the use of time to facilitate student learning.
6	Uses a variety of resources and materials.
7	Provides learning experiences which enable students to transfer principles and generalizations to situations outside of school.
8	Provides group communication (cooperation, interaction, learning from others) experiences for students.
9	Uses a variety of functional verbal and non-verbal communication skills with students.
10	Gives clear directions and explanations.
11	Motivates students to ask questions.
12	Accepts varied student viewpoints and/or asks student to extend or elaborate answers or ideas.
13	Provides feedback to learners on their cognitive performance.
14	Maintains an environment in which students are actively involved, working on-task.
15	Implements an effective classroom management system for positive student behavior (discipline).
16	Uses positive reinforcement patterns with students.
17	Assists students in discovering and correcting errors and inaccuracies.
18	Develops student feedback, evaluation skills and student self-evaluation.

Results

The correlations among the competencies were factored using the SAS procedures for principal components. Table 2 presents the Promax rotated factor pattern for the 18 competencies as measured by the "higher inference" TPAI. There was one very sizable factor, virtually a general factor, and one lesser factor. Such a general factor suggests that the student teachers when measured on the TPAI tend to perform consistently well or poorly across the set of competencies. In other words there is strong correlation across the competencies. Whether this is due to a "halo effect," the natural overlap among the competencies, or some other consideration is not known.

The second set of results is presented in Table 3. These results are for the 18 competencies as measured by the COKER. There were five strong factors among the 18 competencies. The Promax solution which is reported in Table 3 has reasonable good simple structure. It is interesting to note the differences between this solution and the solution in Table 2. There is no large, general factor here. Rather, the percentages of variance attributable to the factor are roughly equal across the factors. So, although about the same percentage of the variance is accounted for by each solution, the patterns of these factors are markedly different.

It is relatively straightforward to describe the two factors from the TPAI solution. The first factor is a general teaching factor since thirteen of the eighteen competencies load highly on that factor. The second factor has competencies 1, 3, 6, 13 and 18 with high loadings. These competencies are concerned with instructional variety and feedback to students.

The COKER factors are also quite clear. The first factor is concerned with variety and flexibility. The second deals with modifying instruction and working with various sized groups. Factor three is a discipline and communication factor. The fourth factor is concerned with questioning and discussion

Table 2

Factor Pattern for the TPAI Competencies which are also Measured by the COKER

Competency	Factor I	Factor II
1	.40	.67
2	.93	-.02
3	-.20	.99
4	.63	.33
5	.87	-.09
6	.18	.62
7	.69	.12
8	.79	-.13
9	.82	.18
10	.83	.13
11	.86	.08
12	.64	.19
13	.18	.80
14	.98	-.06
15	.90	.01
16	.91	.05
17	.40	.64
18	-.16	.99
Eigen Value	11.96	2.05
% Variance	11.11	8.10

Table 3

Factor Pattern for the COKER Competencies
which are also Measured by the TPAI

	Factor				
Competency	I	II	III	IV	V
1	.65	.16	-.02	.21	.25
2	.83	.02	.02	.07	.13
3	-.02	.85	.05	.20	-.18
4	-.04	.72	-.17	.12	.08
5	-.90	.03	.07	-.05	.01
6	.30	.43	.08	.17	-.26
7	.21	-.05	.09	.62	.27
8	-.17	.34	.50	-.26	-.17
9	-.19	-.43	.65	.24	-.03
10	.41	-.14	.44	-.59	-.12
11	.46	.27	.09	.59	-.06
12	.21	.01	-.02	.84	-.08
13	-.16	.01	.12	.46	.69
14	-.36	.54	.50	-.03	.16
15	.16	.04	.80	-.02	.14
16	.29	-.13	.35	-.04	.58
17	.33	.03	-.28	-.05	.64
18	.27	.70	.02	-.25	.15
Eigen Value	4.88	2.62	2.06	1.68	1.10
% Variance	3.61	2.92	2.13	2.92	2.27

while the fifth factor deals with feedback and correction of cognitive performance.

Although each solution is meaningful, it is clear that the two solutions are very different. The TPAI produced a very general factor while the COKER produced separate and fairly equal (in terms of variance accounted for) factors.

To investigate whether the same variance was being accounted for in the two solutions, the 36 competencies (two measures per competency for 18 competencies) were factored together. The Promax solution is presented in Table 4. Surprisingly, this result is essentially the combination of the results from the separate analyses that are reported in Tables 2 and 3. Factors I and VI are the TPAI factors from Table 2 and factors II, III, IV, V, and VII are the COKER factors from Table 3.

The anticipated merging of the same competencies onto the same factors did not occur. The factor structure in this case was determined more by the method of measurement than by the competencies that were measured. This is clearly an unanticipated and intriguing finding. In Campbell-Fiske multitract, multi-method terms, this solution indicates that two traits measured by the same method correlated higher than the same trait measured by two methods. Such a finding then brings into question the construct validity of at least one of the instruments.

The results of this study indicate little convergent validity between the low-inference and high-inference observation. This indicates that measurement in this area is highly, instrument dependent. With the present state of the art, it seems that both types of measures should be maintained. One consequence is that when reporting information on student teachers (or inservice teachers) the results must be qulified as to the type of observation used. Certainly there is a need for continued research on the construct

Factor Structure for Competencies
Measured by COKER and TPAI
(loadings above .40)

		Factor						
		I	II	III	IV	V	VI	VII
COKER	1		.87					
	2		.90					
	3			.45	.78			.43
	4				.70			.44
	5		-.89					
	6							.78
	7					.61		
	8			.78				
	9				-.51			
	10			.44	-.48			
	11		.45		.75			
	12				.66			
	13					.83		
	14			.67				
	15			.67				
	16		.52					
	17		.73					
	18							.64
TPAI	1	.92					.43	
	2	.89						
	3	.63					.78	
	4	.83						
	5	.80						
	6	.60				.43	.50	
	7	.76						
	8	.69						
	9	.92						
	10	.89						
	11	.83						
	12	.69					.43	
	13	.79					.55	
	14	.93						
	15	.90						
	16	.94						
	17	.84					.46	
	18	.60					.81	
Eigen Value		12.8	4.7	2.9	2.4	1.7	1.1	1.1
% Variance		12.38	4.50	2.64	3.35	3.65	3.72	2.04

validity of observation inventories. As is often the case with exploratory research, this study may raise more questions than it answers.

Note: For additional information and a review of the literature on the measurement of teacher performance through observation, the reader is referred to: G. E. Dickson and W. Wiersma, Measurement of Teacher Competence; Toledo, OH., Center for Educational Research and Services, The University of Toledo, 1982.

ABSTRACT

Assessment of Teacher Performance: Constructs of Teacher Competencies Based on Factor Analysis of Observational Data

This study focused on a comparison of the factor structures of the scores on 18 teacher competencies measured by both low-inference and high-inference observation. The observation inventories used were the Classroom Observations Keyed for Effectiveness Research (COKER) for low-inference observation, and the Teacher Performance Assessment Instructions (TPAI) for high-inference observation. The data were obtained on over 250 student teachers observed in their concluding year of training, in a four-year, competency-based, teacher education program.

The factor structures based on the two sets of scores, one from the COKER, the other from the TPAI, were quite different; although variance accounted for was about the same. The TPAI produced a strong, general factor and a second factor concerned with instructional variety. The COKER produced five, consistently strong factors. The results indicated little convergent validity between the low-inference and high-inference observation. Measurement in this area appears to be highly, instrument dependent.